NPS55-85-004

NAVAL POSTGRADUATE SCHOOL

Monterey, California

"PROCESSOR-SHARED TIME-SHARING

MODELS IN HEAVY TRAFFIC"

by

Donald P. Gaver

Patricia A. Jacobs

March 1985

Prepared for;

Office of Naval Research
Arlington, Va 22217

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral R. H. Shumaker
Superintendent

David A. Schrady
Provost

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPS55-85-004
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE *(and Subtitle)*: PROFESSOR-SHARED TIME-SHARING MODELS IN HEAVY TRAFFIC
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Donald P. Gaver, Patricia A. Jacobs
8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, CA 93943
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61153N; RR014-05-01, N0001485WR24061
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Arlington, VA 22217
12. REPORT DATE: March 1985
13. NUMBER OF PAGES: 53
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*:
15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT *(of this Report)*: Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*:
18. SUPPLEMENTARY NOTES:
19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*:
20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*:

Probability models are presented for computer systems with processorshared (time sliced) service discipline. The response (sojourn) time of an arriving job that requires T units of processing time is shown to be approximately Gaussian/normal under moderately heavy traffic conditions, e.g. when the number of terminals becomes large.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

PROCESSOR-SHARED TIME-SHARING MODELS IN HEAVY TRAFFIC

Donald P. Gaver
Patricia A. Jacobs

1. Introduction

Processor sharing (PS) is a mathematically tractable approximation to time sharing, a procedure followed in many actual computer systems. In effect, PS assigns to each job of the i, (i = 1,2,...) present for processing 1/ith of the total processing effort; equivalently, a single job with Markovian service rate μ completes processing in $(t,t+dt)$ with probability $(\mu/i)dt + o(dt)$. One advantage of PS is that short jobs are not trapped behind long jobs, as is possible in a FC-FS discipline.

Various mathematical results have been obtained about certain processor sharing models. An early example was the paper of Coffman et al.(1970). Recently extensive results have been obtained for Markovian systems by D. Mitra (1981), and for non-Markovian single-server Poisson arrival systems by T. Ott (1984) and V. Ramaswami (1984).

This paper is a continuation of work reported in Gaver, Jacobs, and Latouche (1984), henceforth GJL, where emphasis was placed on proposing and evaluating simple approximations to the distribution of delay experienced by a particular "tagged" job approaching a time-shared processor. In that paper it was shown that under heavy traffic conditions, or if the tagged job duration became large, then the distribution of tagged job response time (also called sojourn time by others) approaches the normal or Gaussian distribution.

We first show that the analysis of our system and the problem solution can naturally and conveniently be conducted in work time rather than in ordinary clock time. Subsequently we focus upon heavy-traffic approximations to the distribution of response time, R(T), given the actual work (computer time) requirement, T. Specific models proposed and examined are (i) for a system of many identical terminals that independently submit jobs or programs according to the same Markovian process, and (ii) for a system having two terminal types, each of which submits jobs in a manner governed by its own Markov process. The methodology extends to general k terminal types as well, and to other models.

The approximation solutions are evaluated for accuracy by means of Monte Carlo simulations.

2. The Work-Time Concept

Imagine that a tagged job requiring T units of processing approaches the computer. Assume that it arrives when the system is in steady state. After that initial moment it undergoes processing, at various rates governed by the amount of its accompaniment, until T units of service or work are accumulated, at which point it departs after a random delay of R(T). In following the tagged job's delay, it turns out to be convenient to measure time in terms of the amount of actual work or processing that has been accomplished on the tagged job. Thus let $\{X(w), w \geq 0\}$ denote the number of programs or jobs undergoing service at a moment when exactly w units of processing have been accomplished on the tagged job. The instantaneous rate of accrual of clock or response time at work time w is clearly X(w): if X(w) = 1 then the tagged job is alone and response (clock) time and work time advance at the same rate, while if X(w) = 17 the tagged job is accompanied by 16 others and 17 units of response time accrue for every single work time unit. It follows that

$$R(T) = \int_0^T X(w)\, dw \,. \qquad (2.1)$$

For the models to be considered, the process X(w) is a birth-and-death, or simple Markov, process related to N(t), the number of jobs in the system at clock time t; its transition rates, after adjusting for tagged job entry, are seen to be

$$\lambda_i^t dt = \lambda(N-i)\,dt = \lambda(N-i)i\,dw = \lambda_i^w dw \qquad (2.2)$$

$$\mu_i^t dt = \mu dt = \mu\, i\, dw = \mu_i^w dw$$

for $1 \leq i \leq N$ where N is the total number of terminals in the system. The term idw is required to allow the work-time process to advance appropriately in clock time. Henceforth we drop the superscripts, allowing the context to imply the appropriate rate.

The approach taken here invokes the work time concept described above to facilitate calculations, first for a single terminal or job type situation, but later on for a system with more diversified traffic patterns.

3. Heavy Traffic Analysis of a Single Terminal Type Markovian System

Suppose that N terminals have access to a single computer. Each terminal has Markovian demand rate λ, and expected service time μ^{-1}. Service times are assumed to be independently and exponentially distributed. The discipline at the computer is PS. It is clear that if N(t) is the number of terminals that have submitted jobs that are undergoing service at the computer at (clock) time t, then $\{N(t), t \geq 0\}$ is a Markov process in continuous time that is identical to the classical single repairman problem; see Feller (1957), p. 416. This is so, since if $N(t) = i > 0$, then each individual job or program receives (dt/i) units of processing time in $(t,t+dt)$, and hence departs with probability $\mu(dt/i) + o(dt)$, but the probability that some job departs is $i\mu(dt/i) + o(dt) = \mu dt + o(dt)$. It has been shown by Iglehart (1965) and by Burman (1979) that under heavy traffic conditions $(N \to \infty)$ one may approximate N(t) by a suitable Gaussian process, namely the Ornstein-Uhlenbeck process. This fact alone enables one to study the distribution of R(T), and to deduce approximate normality; see GJL for a first analysis.

3.1 Diffusion Approximation in Work Time.

Here is a diffusion process approximation for X(w). On the basis of intuition write down the stochastic differential equation for $\tilde{X}(w)$, the approximation to X(w):

$$d\tilde{X}(w) = \lambda[N-\tilde{X}(w)]\tilde{X}(w)dw - \mu\tilde{X}(w)dw + \sqrt{\lambda[N-\tilde{X}(w)]\tilde{X}(w) + \mu\tilde{X}(w)}\, dB(w) \tag{3.1}$$

where $\{B(w),\ w \geq 0\}$ is standard Brownian motion. The first right-hand-side term represents infinitesimal drift of $\tilde{X}(w)$, while the second is the diffusion or infinitesimal variance term, the form of which is obtained from the observation that arrival and departure processes compete like independent Poisson processes in short time periods. Now suppose that, as $N \to \infty$,

$$\tilde{X}(w) \simeq Nm(w) + \sqrt{N}\, Z(w) \tag{3.2}$$

where $m(w)$ is a deterministic function of time, and $\{Z(w),\ w \geq 0\}$ is a stochastic process, the properties of which must be discovered. Substitute (3.2) into (3.1) to obtain

$$\begin{aligned} N dm(w) + \sqrt{N} dZ(w) &= \lambda [N - Nm(w) - \sqrt{N} Z(w)][Nm(w) + \sqrt{N} Z(w)]dw \\ &\quad - \mu [Nm(w) + \sqrt{N} Z(w)]dw \\ &\quad + \sqrt{\lambda [N - Nm(w) - \sqrt{N} Z(w)][Nm(w) + \sqrt{N} Z(w)] + \mu [Nm(w) + \sqrt{N} Z(w)]}\ dB(w) \ . \end{aligned} \tag{3.3}$$

Next isolate terms of order N and $\sqrt{N}$; the result is, after stipulating that $\lambda' = \lambda N$, a constant as $N \to \infty$,

$$O(N): \quad \frac{dm(w)}{dw} = \lambda'[1 - m(w)]m(w) - \mu m(w) \ , \tag{3.4}$$

$$\begin{aligned} O(\sqrt{N}): \quad dZ(w) &= \{\lambda'[1 - 2m(w)] - \mu\} Z(w) dw \\ &\quad + \sqrt{\lambda'[1 - m(w)]m(w) + \mu m(w)}\ dB(w) \ ; \end{aligned} \tag{3.5}$$

the stochastic differential equation (3.5) is of Ornstein-Uhlenbeck (O-U) form; see Arnold (1974).

Next obtain the approximate long-run mean as the solution of (3.4) with dm/dw = 0, examining only the heavy-traffic situation in which $\lambda' > \mu$:

$$m(\infty) = 1 - \frac{\mu}{\lambda'}, \qquad \lambda' > \mu \tag{3.6}$$

$$\equiv 1 - \frac{\mu}{\lambda N}$$

If the above solution is used to define the stochastic differential equation parameters there results

$$dZ(w) = (-\lambda'+\mu)\,Z(w)\,dw + \sqrt{2\mu(1-(\mu/\lambda'))}\;dB(w), \tag{3.7}$$

which suggests that $\{Z(w)\}$ can be considered an O-U process with constant coefficients; namely

$$dZ(w) = -\rho Z(w)\,dw + \sigma dB(w), \tag{3.8}$$

the solution to which is

$$Z(w) = Z(0)e^{-\rho w} + \sigma \int_0^w e^{-\rho u} dB(u). \tag{3.9}$$

The parameters $\rho = (\lambda' - \mu)$ and $\sigma^2 = 2\mu(1-\mu/\lambda')$.

3.2 Response Time Evaluation

Let

$$\tilde{R}(T) = \int_0^T \tilde{X}(w)\,dw \simeq \int_0^T [Nm(w) + \sqrt{N}Z(w)]\,dw \tag{3.10}$$

approximate the response time; in this approximation $\tilde{R}(T)$ is normally distributed (Gaussian). First,

$$E[R(T)] \simeq E[\tilde{R}(T)] \simeq N \int_0^T m(w)\,dw = N(1 - \tfrac{\mu}{\lambda'})T \qquad (3.11)$$

Second,

$$\mathrm{Var}[R(T)] \simeq \mathrm{Var}[\tilde{R}(T)] = \mathrm{var}[\sqrt{N} \int_0^T Z(w)\,dw] \qquad (3.12)$$

$$= N\{E[\int_0^T Z(w)\,dw \int_0^T Z(u)\,du)] - (E[\int_0^T Z(w)\,dw])^2\};$$

for ease of writing we have left the initial condition $Z(0)$ implicit. In order to evaluate the above, recall that the tagged job approaches the server when the latter is in equilibrium, i.e., at $t = \infty$. It may be shown that the diffusion approximation for $N(t)$, the number undergoing service at clock time t, is

$$\tilde{N}(t) = Na(t) + \sqrt{N}\, Y(t) \qquad (3.13)$$

where $a(t)$ is a deterministic function of time and $\{Y(t)\}$ is a particular Ornstein-Uhlenbeck process. A similar analysis to that leading to (3.6) and (3.7) yields

$$a(\infty) = 1 - \frac{\mu}{N\lambda}, \quad N\lambda > \mu, \qquad (3.14)$$

$$= 1 - \frac{\mu}{\lambda'}$$

and

$$E[Y(\infty)] = 0, \quad Var[Y(\infty)] = \frac{\mu}{N\lambda} \equiv \frac{\mu}{\lambda'} ; \tag{3.15}$$

see GJL for a derivation; (3.13) provides the initial condition for evaluating moments of R(T), using (3.10). Identify Z(0), the initial value of the work time noise process Z(w), with Y(∞). According to (3.9), this implies that $E[\int_0^T Z(w)\,dw] = 0$. In order to compute the Var[R(T)] it is next necessary to evaluate the following integral:

$$\begin{aligned} I(T) &= E[\int_0^T Z(w)\,dw \int_0^T Z(u)\,du] \\ &= E[2\int_0^T Z(w)\,dw \int_w^T E[Z(u)\,|\,Z(w)]\,du] \\ &= 2E[\int_0^T Z(w)\,dw \int_w^T Z(w)\,e^{-\rho(u-w)}du] \qquad (3.16) \\ &= 2E[\int_0^T Z(w)\,dw Z(w)\frac{1}{\rho}(1 - e^{-\rho(T-w)})] \\ &= 2E[\int_0^T E[Z(w)^2\,|\,Z(0)]\frac{1}{\rho}(1 - e^{-\rho(T-w)})dw] \ . \end{aligned}$$

Square (3.9) and take the expectation to see that, condition-ally on Z(0),

$$I(T) = 2E\left[\int_0^T (Z(0)^2 e^{-2\rho w} + \frac{\sigma^2}{2\rho}(1-e^{-2\rho w}))\frac{1}{\rho}(1-e^{-\rho(T-w)})\,dw\right]$$

$$= \frac{2}{\rho}E[Z(0)^2]\left[\int_0^T e^{-2\rho w}(1-e^{-\rho(T-w)})\,dw\right]$$

$$+ \frac{\sigma^2}{\rho^2}\int_0^T (1-e^{-2\rho w})(1-e^{-\rho(T-w)})\,dw\ .$$

Now put $E[Z(0)^2] = E[Y(\infty)^2] = \mu/\lambda' = \sigma^2/2\rho$ to see that

$$I(T) = \frac{\sigma^2}{\rho^2}\int_0^T [1-e^{-\rho(T-w)}]\,dw = \frac{\sigma^2}{\rho^2}[T - \frac{1}{\rho}(1-e^{-\rho T})]$$

$$= \frac{\sigma^2 T}{\rho^2}[1 - \frac{1}{\rho T}(1-e^{-\rho T})]\ .$$

Thus it follows that

$$\mathrm{Var}[R(T)] \sim \mathrm{Var}[\tilde{R}(t)] = NT\,\frac{\sigma^2}{\rho^2}[1 - \frac{1}{\rho T}(1-e^{-\rho T})]\ . \qquad (3.14)$$

To terms of order T this agrees with (4.10) of GJL; not surprisingly the additional factor in (3.14) can actually provide numerical results superior to those of GJL.

The form of the heavy traffic approximation, namely the limiting normal form with parameters (3.11) and (3.14), can be more rigorously validated by use of the theory of convergence of suitably normalized sequences of semigroups of transformations; see Burman (1979). Details appear in an appendix to this paper.

4. Heavy Traffic Analysis of a K-Terminal-Type Processor Sharing System.

Consider the following natural extension of the previous model. The processor is jointly utilized by K sets of terminals, each generating distinctive job types. There are N_i terminals in the ith set, and arrival rate and service rate are λ_i and μ_i respectively. Again the discipline at the computer is PS. Of course this is not the same as a situation in which all terminals are the same, but Type j jobs occur with probability p_j from each terminal. The latter model can, however, be studied in an analogous heavy-traffic manner, as can other interesting models.

4.1 A Diffusion Model for the Work-Time Process.

Let $\{X_i(w),\ i = 1,\ldots,K\}$ represent the number of jobs of all types present at the computer at work time w. The present model implies that $\{X_i(w)\}$ is a multivariate or vector-state birth and death Markov process. We choose to study a diffusion approximation $\{\tilde{X}_i(w)\}$ to $\{X_i(w)\}$ that is described by the following system of s.d.e.:

$$d\tilde{X}_i(w) = \lambda_i(N_i - \tilde{X}_i(w))\left(\sum_{k=1}^{K} \tilde{X}_k(w)\right)dw - \mu_i \tilde{X}_i(w)\,dw$$

$$+ \sqrt{\lambda_i(N_i - \tilde{X}_i(w))\sum_{k=1}^{K} \tilde{X}_k(w) + \mu_i \tilde{X}_i(w)}\ dB_i(w) \qquad (4.1)$$

$$i = 1,2,\ldots,K$$

where $\{B_i(w)\}$ are mutually independent standard Brownian motion

or Wiener processes. The work time process is a transformation of the clock time process; in particular, the drift of the ith component of the clock time process $\{N_i(t)\}$ is seen to be

$$\lambda_i(N_i - N_i(t))dt - \mu_i \frac{N_i(t)dt}{\sum_{k=1}^{K} N_k(t)}, \tag{4.2}$$

which exhibits the processor-sharing effect in the term multiplying μ_i. Multiplication by the total in service, $\sum N_i(t)$, converts to the work time transition rates, in analogy with (2.2).

Now once again approximate by writing

$$\tilde{X}_i(w) = N_i m_i(w) + \sqrt{N_i}\, Z_i(w), \quad i = 1,2,\ldots,K; \tag{4.3}$$

$m_i(w)$ and $\{Z_i(w)\}$ are to be determined, subject to the normalization $N = \sum_{k=1}^{K} N_k \to \infty$ but with

$$N_i/N \to \ell_i, \quad 0 \leq \ell_i \leq 1, \tag{4.4a}$$

and

$$N\lambda_i \to \lambda_i', \quad 0 \leq \lambda_i' \leq \infty, \quad \lambda_i' > \mu_i. \tag{4.4b}$$

Conditions (4.3) and (4.4) are referred to as the heavy traffic normalization (HTN). The result of isolating terms according to order in (4.1) is:

$$O(N): \quad \frac{dm_i(w)}{dw} = \lambda_i'(1-m_i(w)) \sum_{k=1}^{K} \ell_k m_k(w) - \mu_i m_i(w) \qquad (4.5)$$

$$O(\sqrt{N}): \quad dZ_i(w) = \lambda_i'\{-(\sum_{k=1}^{K} \ell_k m_k(w))Z_i(w) + \sqrt{\ell_i}(1-m_i(w)) \sum_{k=1}^{K} \sqrt{\ell_k} Z_k(w)\}dw$$

$$- \mu_i Z_i(w)dw \qquad (4.6)$$

$$+ \sqrt{\lambda_i'(1-m_i(w)) \sum_{k=1}^{K} \ell_k m_k(w) + \mu_i m_i(w)}\ dB_i(w)$$

for $i = 1,2,\ldots,K$. Thus $(m_i(w);\ w \geq 0;\ i = 1,2,\ldots,K)$ must be found by solving a system of ordinary first-order, but non-linear differential equations, while (4.6) shows that $\{Z_i(w);\ w \geq 0,\ i = 1,2,\ldots,K\}$ is a multivariate Ornstein-Uhlenbeck process.

4.2 A Diffusion Model for the Clock Time Process.

In order to provide the initial conditions encountered by the tagged job, it is necessary to study the clock-time process $N_i(t)$; see (3.13). The corresponding approximation has s.d.e.

$$d\tilde{N}_i(t) = \lambda_i(N_i - \tilde{N}_i(t))dt - \mu_i \frac{\tilde{N}_i(t)}{\sum_{k=1}^{K} \tilde{N}_k(t)}\, dt$$

$$+ \sqrt{\lambda_i(N_i - \tilde{N}_i(t)) + \mu_i \frac{\tilde{N}_i(t)}{\sum_{k=1}^{K} \tilde{N}_k(t)}}\ dB_i(t) \quad i = 1,2,\ldots,K\ . \qquad (4.7)$$

Now invoke the HTN:

$$\tilde{N}_i(t) = N_i a_i(t) + \sqrt{N_i}\, Y_i(t) , \tag{4.8}$$

and again $N = \sum_{k=1}^{K} N_k \to \infty$, with

$$N_i/N \to \ell_i , \quad 0 \le \ell_i \le 1 , \tag{4.9}$$

but

$$\mu_i/N \to \mu_i' , \quad 0 \le \mu_i' < \infty, \quad \lambda_i > \mu_i' . \tag{4.10}$$

The result of isolating terms is

$$O(N): \quad \frac{da_i(t)}{dt} = \lambda_i(1 - a_i(t)) - \mu_i' \frac{a_i(t)}{\sum_{k=1}^{K} \ell_k a_k(t)} \tag{4.11}$$

$$O(\sqrt{N}): \quad dY_i(t) = -\lambda_i Y_i(t)\,dt - \mu_i' \left\{ \frac{Y_i(t)}{\sum_{k=1}^{K} \ell_k a_k(t)} + \sqrt{\ell_i}\, a_i(t) \frac{\sum_{k=1}^{K} \sqrt{\ell_k} Y_k(t)}{\left(\sum_{k=1}^{K} \ell_k a_k(t)\right)^2} \right\} dt$$

$$+ \sqrt{\lambda_1(1-a_i(t)) + \mu_i' \frac{a_i(t)}{\sum_{k=1}^{K} \ell_k a_k(t)}}\; dB_i(t) . \tag{4.12}$$

These equations closely resemble those describing the work time approximation; again the semigroup approach is applicable.

If a long-run solution to the O(N) term exists in work time, and consequently $dm_i/dw \to 0$ as $w \to \infty$, the result is the system of equations for $m_i(\infty) \equiv m_i$:

$$\rho_i(1 - m_i) \sum_{k=1}^{K} \ell_k m_k - m_i = 0 , \tag{4.13}$$

where $\rho_i = \lambda_i'/\mu_i = N\lambda_i/\mu_i$. Now these same equations are satisfied by a presumed long-run solution in clock time, i.e., if $da_i/dt \to 0$ in (4.11); for $a_i(\infty) = a_i$:

$$\rho_i(1 - a_i) - a_i / \sum_{k=1}^{K} \ell_k a_k = 0 . \tag{4.14}$$

Consequently the long-run solutions in work and clock time agree at the O(N) term level; this means that the long-run mean number present in both clock and work time agree:

$$E[N_i(t)] \sim Na_i(\infty) = Nm_i(\infty) \sim E[X_i(t)] . \tag{4.15}$$

Next substitute these long-run results in the s.d.e. to see that as $t,w \to \infty$, Y_i and Z_i are essentially the same process. Put $S = \sum_{k=1}^{K} \ell_k a_k$ to simplify writing. Then

$$dY_i(t) = -\lambda_i Y_i(t)\,dt - \mu_i' \frac{Y_i(t)}{S} dt + \lambda_i(1-a_i)\sqrt{\ell_i}\, \frac{\sum_{k=1}^{K} \sqrt{\ell_k} Y_k(t)}{S}\, dt$$

$$+ \sqrt{2\lambda_i(1-a_i)}\, dB_i , \tag{4.16}$$

or

$$dY_i(t) = \lambda_i\{-Y_i(\varepsilon)S + \sqrt{\ell_i}(1-a_i)\sum_{k=1}^{K}\sqrt{\ell_k}Y_k(t)\}\frac{dt}{S} - \mu_i' Y_i(t)\ \frac{dt}{S}$$

$$+ \sqrt{2\lambda_i(1-a_i)S}\ \frac{1}{\sqrt{S}}\ dB_i(t)\ , \qquad (4.17)$$

and a direct comparison with the corresponding equation in work time, (4.6), shows that the long-run behaviors of the two processes $\{Z_i(w)\}$ and $\{Y_i(t)\}$ are identical except for a constant time-scale change: for large w and t,

$$\{Z_i(w)\} \quad \text{and} \quad \{Y_i(\tfrac{t}{NS})\} \qquad (4.18)$$

have the same probability law; i.e., finite-dimensional distributions and limiting distribution.

4.3 Response Time

We discuss the response time under these conditions: a tagged job approaches the processor when the latter has been operating for some time, so the long-run clock time distribution prevails; after arrival, the job remains present until the total work time accumulated on the job is T, the requested service time, giving

$$R(T) = \int_0^T \sum_{i=1}^K [X_i(w)\,dw] \simeq \sum_{i=1}^K \int_0^T \tilde{X}_i(w)\,dw \qquad (4.19)$$

$$= N \sum_{i=1}^K \ell_i \int_0^T m_i(w)\,dw + \sqrt{N} \sum_{i=1}^K \sqrt{\ell_i} \int_0^T Z_i(w)\,dw$$

where it is understood that the initial condition for the $Z_i(w)$ integrand in (4.19) is given by the approximate stationary distribution from the clock time process. In view of (4.18), this is equivalent to removing the initial condition by the long-run distribution of the work time process itself.

Since the long-run situation is being discussed it is first necessary to solve the steady-state version of (4.5)

$$0 = \lambda_i'(1 - m_i) \sum_{k=1}^K \ell_k m_k - \mu_i m_i, \quad i = 1,2,\ldots,K. \qquad (4.20)$$

Then the solution provides parameters for the long-run version of (4.6); here written in matrix form

$$d\underline{Z}(w) = \underline{A}\,\underline{Z}(w)dw + \underline{\sigma}\,d\underline{B} \qquad (4.21)$$

Now to find the variance of R(T), append the row

$$dZ_{k+1}(w) = \sum_{i=1}^K \sqrt{\ell_i}\, Z_i(w)dw \qquad (4.22)$$

to the former drift matrix $\underline{A}$ of (4.21), and consider the system

$$d\underline{Z}^*(w) = \underline{A}^*\underline{Z}^*(w)dw + \underline{\sigma}^*d\underline{B}^* \qquad (4.23)$$

the solution to which can be formally written out in terms of the appropriate fundamental matrix, and computed in terms of eigenvalues and eigenvectors of the matrix $\underline{A}^*$. See e.g. Arnold (1974), Chapter 8 and Coddington and Levinson (1955) for details. A convenient way of formalizing the calculations is actually by using Laplace transforms. Unfortunately, no truly simple formulas result. Finally, the covariance matrix $\underline{C}(w)$ of the components of $\underline{Z}^*(w)$ satisfies the matrix differential equation

$$\frac{d\underline{C}(w)}{dw} = (\underline{A}^*)\underline{C}(w) + \underline{C}(w)(\underline{A}^*)' + (\underline{\sigma})(\underline{\sigma})'$$

where ' denotes transpose; the initial conditions are provided by the long-run distribution in clock time, or in view of (4.18), of the work time process $\{\underline{Z}\}$ itself. It is the $K+1^{\underline{st}}$ diagonal element of $\underline{C}(w)$, evaluated at $w = T$ and multiplied by N that provides the required approximate Var[R(T)].

5. Simulation Studies of the Accuracy of the Normal Approximations to the Distribution of Response Time

In this section we use simulation to study the numerical accuracy of normal approximations to the distribution of the response time. Two continuous time Markov chain models were simulated. In one there is a single terminal type; in the second, a two-terminal type system is examined. Two normal approximations were evaluated: one results from a central limit theorem, and the other results from applying the previously derived diffusion approximation to the Markov processes.

5.1 A Single Terminal Type Markovian System.

Let $\bar{X}(w)$ denote the number of other jobs undergoing service at a moment when exactly w units of processing has been accomplished on the tagged job for the single terminal type Markovian model of Section 3. Since $\{\bar{X}(w);\ w \geq 0\}$ is a Markov process and

$$R(T) = \int_0^T (\bar{X}(w) + 1)\ dw\ ,$$

it follows that there are constants $m(c)$ and $\sigma(c)$ such that

$$\frac{R(T) - m(c)T}{\sigma(c)\sqrt{T}}$$

converges in distribution to a standard normal distribution as $T \to \infty$ (cf. Keilson (1979), p. 121). Call this a central limit theorem (CLT) for such a process. In this case

$$m(c) = 1 + \sum \pi(j)j$$

where π is the stationary distribution of $\{\bar{X}(w);\ w \geq 0\}$ and a formula for evaluating $\sigma(c)$ is given in Keilson (1979). The CLT normal approximation states that R(T) has a normal distribution with mean $m(c)T$ and variance $\sigma(c)^2T$. The CLT mean is the true mean for R(T) under steady state (cf. GJL). The CLT normal approximation should be increasingly accurate as T becomes large, despite values of other system parameters, including the number of terminals.

The derivation of the heavy traffic (or diffusion) approximation is detailed in Section 3. In summary, the HT approximation is that R(T) has a normal distribution with mean

$$N(1 - \frac{\mu}{\lambda N})T , \qquad \frac{\mu}{\lambda N} < 1$$

and variance

$$NT \frac{\sigma^2}{\rho^2}[1 - \frac{1}{\rho T}(1 - e^{-\rho T})] ,$$

where

$$\sigma^2 = 2\mu(1 - \frac{\mu}{N\lambda})$$

and

$$\rho = N\lambda + \mu .$$

The mean and variance of the HT approximation are easier to

compute than those for the CLT. It is anticipated that the HT approximation should be increasingly accurate as N becomes large when heavy traffic conditions prevail, i.e. $\frac{\mu}{\lambda N} < 1$. It is inapplicable under other circumstances. We have conducted simulations to assess these anticipations. All simulations were carried out on an IBM 3033 computer at the Naval Postgraduate School using the LLRANDOMII random number generating package (see Lewis and Uribe (1981)).

Conditional response times given the number of jobs being processed at the time of arrival of the tagged job were simulated; the tagged job required T time units of processing. For each initial condition, 500 replications were done. Sample moments and relative frequencies were computed for each initial condition giving conditional response time sample moments, and selected response time relative frequencies, i.e., estimated probabilities of response times in selected ranges. Unconditional sample moments and relative frequencies were then computed by multiplying each conditional moment or relative frequency by the appropriate stationary probability of there being j jobs present at the time of arrival of the tagged job and then summing over all possible j. The stationary probability is of the form $k\lambda\pi(j)$ where k is chosen so that the probabilities sum to 1 (cf. Kelly (1979)). A detailed description of the simulation program can be found in Pornsuriya [1984].

Simulated and approximating means and standard deviations for various values of N, λ, and μ appear in Tables 1-3. Some discussion of specific cases now follows. When N = 10, λ = 25

and $\mu = 100$, it follows from (3.6) that approximately $(10)(1 - \frac{100}{250}) = 6$ jobs are being processed along with the tagged job; thus the traffic is moderate in this case. When $N = 10$, $\lambda = 15$, $\mu = 100$, then on the average 3.3 are being processed with the tagged job; a rather light traffic case. When $N = 25$, $\lambda = 10$, $\mu = 100$, then on the average 15 jobs are being processed; again a moderate traffic case. The HT mean is lower than the simulated mean for $N = 10$. For $N = 25$ it agrees with the simulated mean. As mentioned before, the CLT mean equals the true mean. The CLT standard deviation approaches the simulation value as T becomes large, as anticipated. Also as anticipated, the HT standard deviation is closer to the simulated one for the moderate traffic cases than for the light traffic case. In order to assess the degree of normality of the distribution of R(T), the α-quantiles for each approximating normal distribution were computed. The relative frequency of being less than or equal to each α-quantile was computed using the simulated data. The results appear in Tables 4-6. For $N = 25$, $\lambda = 10$, $\mu = 100$, the HT approximation appears to describe the data well for all values of T. For $N = 10$, the HT approximation does better than the CLT for the moderate traffic case of $\lambda = 25$. For the light traffic case, $N = 10$, $\lambda = 15$, both approximations do poorly for small $T = 0.01$, the mean work request time. The CLT does well for large $T = 0.10$, which is 10 times the average service time.

Note, however, that all simulations have been carried out for modest system sizes, N. If N grows to say 50, or 100, the HT approximations can be expected to improve correspondingly; they are often not bad even at the level of $N = 25$.

TABLE 1

Simulated Mean and Standard Deviation for R(T)
and Their Approximating Values

$N = 10, \lambda = 15, \mu = 100$

TIME T		Mean	Std. Dev.
0.01	Simulation	.0404 (.0002)*	.0172
	CLT	.0404	.0323
	HT	.0333	.0238
0.025	Simulation	.1005 (.0005)	.0375
	CLT	.1010	.0510
	HT	.0833	.0535
0.05	Simulation	.2016 (.0010)	.0622
	CLT	.2019	.0722
	HT	.1667	.0919
0.10	Simulation	.4029 (.0015)	.0940
	CLT	.4039	.1020
	HT	.3333	.1462

* The standard error of the estimate of the mean

TABLE 2

Simulated Mean and Standard Deviation for R(T)
and Their Approximating Values

$N = 10, \lambda = 25, \mu = 100$

TIME T		Mean	Std. Dev.
0.01	Simulation	.0606 (.0002)*	.0503
	CLT	.0605	.0245
	HT	.0600	.0160
0.025	Simulation	.1507 (.0004)	.0315
	CLT	.1513	.0288
	HT	.1500	.0314
0.05	Simulation	.3021 (.0008)	.0497
	CLT	.3027	.0548
	HT	.3000	.0481
0.10	Simulation	.6036 (.0012)	.0748
	CLT	.6053	.0776
	HT	.6000	.0706

*Standard Error for the Estimate of the mean

TABLE 3

Simulated Mean and Standard Deviation for R(T)
and Their Approximating Values

$N = 25, \lambda = 10, \mu = 100$

TIME T		Mean	Std. Dev.
0.01	Simulation	.1498 (.0002)*	.0812
	CLT	.1500	.0390
	HT	.1500	.0254
0.025	Simulation	.3759 (.0006)	.0503
	CLT	.3750	.0617
	HT	.3750	.0497
0.05	Simulation	.7506 (.0010)	.0804
	CLT	.7500	.0872
	HT	.7500	.0760
0.0625	Simulation	.9381 (.0011)	.0909
	CLT	.9375	.0975
	HT	.9375	.0863

*The standard error of the estimate of the mean

TABLE 4

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantile

$N = 10, \lambda = 15, \mu = 100$

TIME	α:	.10	.25	.50	.75	.90	.95	.99
0.01	CLT	0.00	.12	.50	.89	1.0	1.0	1.0
	HT	0.00	.10	.36	.68	.91	.97	1.0
0.025	CLT	0.04	.21	.49	.81	.97	.99	1.0
	HT	0.00	.09	.33	.67	.91	.98	1.0
0.05	CLT	0.08	.23	.49	.77	.94	.98	1.0
	HT	0.0	.07	.30	.65	.92	.97	1.0
0.10	CLT	0.10	.24	.48	.76	.92	.97	1.0
	HT	0.00	.04	.24	.60	.89	.97	1.0

TABLE 5

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantile

$N = 10, \lambda = 25, \mu = 100$

TIME	α:	.10	.25	.50	.75	.90	.95	.99
0.01	CLT	.04	.15	.45	.85	1.0	1.0	1.0
	HT	.11	.22	.43	.72	.91	.98	1.0
0.025	CLT	.08	.19	.45	.80	.98	1.0	1.0
	HT	.10	.22	.44	.73	.92	.98	1.0
0.05	CLT	.09	.22	.45	.77	.95	.99	1.0
	HT	.11	.22	.44	.71	.91	.97	1.0
0.10	CLT	.10	.24	.47	.76	.93	.98	1.0
	HT	.11	.23	.45	.70	.89	.95	1.0

TABLE 6

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantile

$N = 25, \lambda = 10, \mu = 100$

TIME	α:	.10	.25	.50	.75	.90	.95	.99
0.01	CLT	.03	.15	.46	.86	.99	1.0	1.0
	HT	.11	.23	.46	.73	.91	.97	1.0
0.025	CLT	.07	.19	.46	.80	.96	.99	1.0
	HT	.11	.23	.46	.73	.91	.97	1.0
0.05	CLT	.09	.21	.46	.78	.94	.98	1.0
	HT	.11	.24	.46	.73	.91	.96	1.0
0.0625	CLT	.09	.22	.47	.77	.93	.98	1.0
	HT	.11	.24	.47	.73	.90	.96	1.0

5.2 Simulation Results for Markovian Model with Two-Terminal Types.

In this subsection we describe the results of a simulation of the general K-type Markovian model of Section 4, in the case of K = 2 sets of terminals. As in Section 4 let $\bar{X}_i(w)$ represent the number of other jobs of type i being processed when the tagged job has acquired exactly w units of processing. As before the response time for the tagged job requiring T units of work is

$$R(T) = \int_0^T [\bar{X}_1(w) + \bar{X}_2(w) + 1]\, dw .$$

The process $\{(\bar{X}_1(w), \bar{X}_2(w)); w \geq 0\}$ is Markovian. Hence again R(T) satisfies a central limit theorem as $T \to \infty$. The normal approximation for the distribution of R(T) resulting from the central limit theorem will again be referred to as CLT.

The mean term $(m_1 + m_2)T$ for the heavy traffic approximation was computed by solving the system of equations (4.20) for m_1 and m_2. The variance term for the approximation was computed by solving the system of stochastic differential equations (4.23) as detailed in Arnold [Corollary (8.2.4)]. The fundamental matrix (Arnold [p. 129]) was found by computing Laplace transforms of the system of defining differential equations and then inverting the solution. The approximating variance term was found by computing the variance of the solution of the s.d.e. As in the case of one-terminal type, it is a

linear combination of exponentials and constant terms. Its exact form is uninformative and will not be given here.

Conditional response times, given the number of jobs of each type being processed at the time of arrival of the tagged job, were simulated. The tagged job was always taken to be a Type 1 job. For each initial condition, 300 applications were carried out. Sample moments and probabilities (relative frequencies) were computed for each initial condition giving conditional sample moments and probabilities (relative frequencies). Unconditional sample moments and probabilities were computed in a similar manner to that in the one-terminal type simulation; see Pornsuriya [1984].

Values of the simulated means and standard deviations and their approximating values for R(T) for various cases in which $N_1 = 5$ and $N_2 = 5$ appear in Tables 7-10. Once again the CLT mean is the true steady state mean for R(T). The means and standard deviations of R(T) for each T differ surprisingly little for the four cases. This suggests that perhaps the two-type terminal model can often be satisfactorily approximated by a one-type model in which the arrival rate and service rates are the average arrival and service rates in the two-type model. Values of the simulated means and standard deviations and their approximating values for the approximate one-type model with $N = 10$, $\lambda = 25$ and $\mu = 75$ appear in Table 11. The values for the approximate one-type model are acceptably close to those for the two-type model. Note that the quality

of the HT approximation is generally quite good, even though the system sizes N_1 and N_2 can hardly be called "large."

TABLE 7

Simulated Means and Standard Deviations for R(T)
and Their Approximating Values

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 20$, $\lambda_2 = 30$, $\mu_1 = 50$, $\mu_2 = 100$

TIME T		Mean	Std. Dev.
0.01	Simulation	0.0713 (0.0001)*	0.0135
	CLT	0.0707	0.0204
	HT	0.0710	0.0132
0.025	Simulation	0.1783 (0.0003)	0.0263
	CLT	0.1766	0.0322
	HT	0.1775	0.0253
0.0375	Simulation	0.2664 (0.0005)	0.0353
	CLT	0.2650	0.0395
	HT	0.2663	0.0325
0.050	Simulation	0.3570 (0.0005)	0.0414
	CLT	0.3533	0.0456
	HT	0.3550	0.0384
0.0625	Simulation	0.4439 (0.0006)	0.0478
	CLT	0.4416	0.0510
	HT	0.4438	0.0435

*Standard error of the estimate of the mean

TABLE 8

Simulated Means and Standard Deviations for R(T) and Their Approximating Values

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 30$, $\lambda_2 = 20$, $\mu_1 = 100$, $\mu_2 = 50$

TIME T		Mean	Std. Dev.
0.01	Simulation	0.0710 (0.0001)*	0.0137
	CLT	0.0716	0.0204
	HT	0.0710	0.0132
0.025	Simulation	0.1777 (0.0003)	0.0268
	CLT	0.1789	0.0322
	HT	0.1775	0.0253
0.0375	Simulation	0.2672 (0.0004)	0.0351
	CLT	0.2684	0.0395
	HT	0.2663	0.0325
0.050	Simulation	0.3567 (0.0005)	0.0419
	CLT	0.3578	0.0456
	HT	0.3550	0.0384
0.0625	Simulation	0.4461 (0.0006)	0.0476
	CLT	0.4473	0.0510
	HT	0.4438	0.0435

*Standard error of the estimate of the mean

TABLE 9

Simulated Means and Standard Deviations for R(T)
and Their Approximating Values

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 10$, $\lambda_2 = 40$, $\mu_1 = 25$, $\mu_2 = 125$

TIME T		Mean	Std. Dev.
0.01	Simulation	0.0717 (0.0001)*	0.0134
	CLT	0.0717	0.0237
	HT	0.0720	0.0133
0.025	Simulation	0.1788 (0.0004)	0.0280
	CLT	0.1793	0.0375
	HT	0.1799	0.0271
0.0375	Simulation	0.2690 (0.0005)	0.0376
	CLT	0.2684	0.0459
	HT	0.2699	0.0359
0.0500	Simulation	0.3588 (0.0006)	0.0456
	CLT	0.3585	0.0530
	HT	0.3599	0.0433
0.0625	Simulation	0.4481 (0.0007)	0.0527
	CLT	0.4481	0.0529
	HT	0.4498	0.0496

*Standard error of the estimate of the mean

TABLE 10

Simulated Means and Standard Deviations for R(T)
and Their Approximating Values

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 40$, $\lambda_2 = 10$, $\mu_1 = 125$, $\mu_2 = 25$

TIME T		Mean	Std. Dev.
0.01	Simulation	0.0725 (0.0001)*	0.0138
	CLT	0.0724	0.0249
	HT	0.0720	0.0133
0.025	Simulation	0.1816 (0.0004)	0.0290
	CLT	0.1810	0.0394
	HT	0.1799	0.0271
0.0375	Simulation	0.2717 (0.0005)	0.0389
	CLT	0.2714	0.0482
	HT	0.2699	0.0359
0.0500	Simulation	0.3622 (0.0006)	0.0472
	CLT	0.3619	0.0557
	HT	0.3599	0.0433
0.0625	Simulation	0.4522 (0.0008)	0.0557
	CLT	0.4524	0.0623
	HT	0.4498	0.0496

*Standard error of the estimate of the mean

TABLE 11

Simulated Means and Standard Deviations for R(T) and Their Approximating Values for the One-Type Model

$N = 10, \lambda = 25, \mu = 75$

TIME T		Mean	Std. Dev.
0.1	Simulation	0.0701 (0.0002)*	0.0139
	CLT	0.0701	0.0206
	HT	0.0700	0.0135
0.025	Simulation	0.1766 (0.0005)	0.0263
	CLT	0.1752	0.0325
	HT	0.1750	0.0258
0.0375	Simulation	0.2620 (0.0008)	0.0361
	CLT	0.2628	0.0398
	HT	0.2625	0.0330
0.0500	Simulation	0.3501 (0.0009)	0.0430
	CLT	0.3504	0.0460
	HT	0.3500	0.0390
0.0625	Simulation	0.4388 (0.0011)	0.0478
	CLT	0.4380	0.0514
	HT	0.4375	0.0441

*Standard error of the estimate of the mean

To assess the quality of the normal approximation to the distribution of R(T) for the two-type model, the α-quantiles for each two-type approximating normal distribution were computed. The relative frequency of being less than or equal to each α-quantile was then computed, using the simulated data. The results appear in Tables 12-15. From the heavy traffic approximation to the mean it follows that approximately 7 jobs are being processed with the tagged job. Thus, all the cases considered are really moderate traffic cases. The tables indicated that the HT approximation tends to describe the quantiles better than does the CLT. However, as is expected, the CLT improves with larger T.

TABLE 12

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantiles

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 20$, $\lambda_2 = 30$, $\mu_1 = 50$, $\mu_2 = 100$

TIME T	α:	0.10	0.25	0.50	0.75	0.90	0.95	0.99
0.010	CLT	0.041	0.137	0.433	0.838	0.998	1.0	1.0
	HT	0.109	0.221	0.443	0.719	0.915	0.977	1.0
0.0250	CLT	0.062	0.170	0.425	0.765	0.969	0.997	1.0
	HT	0.103	0.223	0.438	0.708	0.911	0.975	1.0
0.0375	CLT	0.079	0.191	0.429	0.754	0.952	0.991	1.0
	HT	0.118	0.229	0.447	0.713	0.905	0.969	0.999
0.0500	CLT	0.075	0.181	0.417	0.724	0.939	0.986	1.0
	HT	0.106	0.222	0.431	0.693	0.897	0.965	0.998
0.0625	CLT	0.081	0.198	0.436	0.745	0.932	0.980	1.0
	HT	0.115	0.238	0.451	0.719	0.898	0.960	0.997

TABLE 13

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantiles

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 30$, $\lambda_2 = 20$, $\mu_1 = 100$, $\mu_2 = 50$

TIME T	α:	0.10	0.25	0.50	0.75	0.90	0.95	0.99
0.0100	CLT	0.049	0.159	0.457	0.869	0.998	1.0	1.0
	HT	0.114	0.229	0.442	0.718	0.919	0.979	1.0
0.0250	CLT	0.078	0.194	0.462	0.809	0.978	0.997	1.0
	HT	0.113	0.228	0.442	0.717	0.916	0.973	0.999
0.0375	CLT	0.088	0.204	0.455	0.787	0.962	0.995	1.0
	HT	0.111	0.224	0.432	0.702	0.909	0.965	0.999
0.0500	CLT	0.088	0.212	0.459	0.771	0.955	0.989	1.0
	HT	0.111	0.226	0.434	0.696	0.898	0.961	0.998
0.0625	CLT	0.093	0.212	0.465	0.773	0.945	0.986	1.0
	HT	0.110	0.222	0.432	0.705	0.889	0.954	0.997

TABLE 14

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantiles

$N_1 = 5,\ N_2 = 5,\ \lambda_1 = 10,\ \lambda_2 = 40,\ \mu_1 = 25,\ \mu_2 = 125$

TIME T	α	0.10	0.25	0.50	0.75	0.90	0.95	0.99
0.010	CLT	0.027	0.119	0.447	0.906	1.0	1.0	1.0
	HT	0.110	0.230	0.458	0.737	0.931	0.982	1.0
0.0250	CLT	0.061	0.178	0.451	0.824	0.991	0.999	1.0
	HT	0.121	0.236	0.464	0.734	0.930	0.985	1.0
0.0375	CLT	0.068	0.191	0.441	0.790	0.979	0.999	1.0
	HT	0.117	0.239	0.451	0.725	0.927	0.983	1.0
0.0500	CLT	0.081	0.189	0.430	0.776	0.975	0.997	1.0
	HT	0.117	0.232	0.444	0.729	0.935	0.985	1.0
0.0625	CLT	0.085	0.199	0.442	0.772	0.960	0.995	1.0
	HT	0.121	0.237	0.456	0.737	0.923	0.979	1.0

TABLE 15

Simulated Probability (Relative Frequency) that the Response Time is Less than or Equal to the Approximating α-Quantiles

$N_1 = 5$, $N_2 = 5$, $\lambda_1 = 40$, $\lambda_2 = 10$, $\mu_1 = 125$, $\mu_2 = 25$

TIME T	α:	0.10	0.25	0.50	0.75	0.90	0.95	0.99
0.0100	CLT	0.024	0.116	0.445	0.907	1.0	1.0	1.0
	HT	0.110	0.228	0.433	0.710	0.905	0.973	1.0
0.0250	CLT	0.054	0.164	0.435	0.820	0.993	1.0	1.0
	HT	0.112	0.219	0.422	0.682	0.901	0.967	1.0
0.0375	CLT	0.070	0.186	0.428	0.795	0.981	0.999	1.0
	HT	0.113	0.224	0.414	0.693	0.899	0.967	0.999
0.050	CLT	0.075	0.182	0.439	0.784	0.972	0.998	1.0
	HT	0.110	0.216	0.422	0.689	0.896	0.969	1.0
0.0625	CLT	0.083	0.200	0.443	0.765	0.959	0.997	1.0
	HT	0.118	0.222	0.427	0.683	0.891	0.958	1.0

APPENDIX

HEAVY TRAFFIC APPROXIMATION BY CONVERGENCE-OF-SEMIGROUPS METHODOLOGY

The purpose of this appendix is to outline a mathematical framework upon which the heavy traffic approximations of this paper may be rigorously based. The approach is to use an analytical theory of convergence of semigroups of operators apparently first applied to queueing problems by Burman (1979) in a regrettably unpublished thesis. See also Lehoczky and Gaver (1981) where the technique is used to obtain results concerning a data-voice traffic sharing multichannel system. The theory of semigroups of operators is introduced in Feller (1971), and detailed in Dynkin (1965); the convergence ideas are discussed in Trotter (1974) and Kato (1976). The basic notion is that the state variable of a process, say the work time process of Section 4.1, $\{X_i(w;N), w \geq 0\}$, is one of a sequence of birth and death Markov processes indexed by system size N. Given such a sequence of Markov processes, $\langle \{X_i(w;N)\} \rangle$, each with its appropriate state space, S_N, it is desired to show that the corresponding sequence of probability transition functions converges to that of some limiting process that has state space S_∞; in the present case $S_\infty = \mathbb{R}_+^K$. The limiting process under the normalization of $\{X(w;N)\}$ chosen will be a particular diffusion process, namely, in the present heavy traffic situation the multivariate Ornstein-Uhlenbeck.

The Trotter-Kato theory of convergence deals with the convergence of infinitesimal operators A_N of the normalized

processes $\{X_i(w);N\}$: if $A_N f \to A_\infty f$ in sup norm for a suitable class of test functions (e.g. $f(z)$: $\mathbb{R}^K \to \mathbb{R}_1$ m-times continuously differentiable, $m \geq 3$, that vanish identically outside a bounded subset of $\mathbb{R}^K$ and further such that the functions, f, together with their first and second derivatives do not increase faster than some fixed power of z) it can be concluded that the semigroups converge, and hence the Markov probability transition functions themselves converge.

We now proceed with the formal calculation of the limiting generator for our normalized process. Invoke (4.4) so

$$Z_i^N(w) = \frac{X_i^N(w) - N_i m_i(w)}{\sqrt{N_i}} = \frac{X_i^N(w) - N\ell_i m_i(w)}{\sqrt{\ell_i}\sqrt{N}} . \qquad (A\text{-}1)$$

By definition, for $z = (z_1, z_2, \ldots, z_K)$ and f in the above class

$$A_N f(\underline{z}) = \lim_{\Delta \to 0} \{E[f(\underline{Z}^N(w+\Delta)) \mid \underline{Z}^N(w) = \underline{z}] - f(\underline{z})\}\frac{1}{\Delta} . \qquad (A\text{-}2)$$

Given $Z_i^N(w) = z_i$, and $C_i(w,w+\Delta)$ represents the change in $X_i^N(w)$,

$$Z_i^N(w+\Delta) = \frac{X_i^N(w) + C_i(w,w+\Delta) - N_i m_i(w+\Delta)}{\sqrt{N_i}}$$

$$= \frac{C_i(w)}{\sqrt{N_i}} + z_i - \sqrt{N_i}\, m_i'(w)\Delta + o(\Delta) . \qquad (A\text{-}3)$$

Consequently, for $\underline{z}$ such that $z_i > -\sqrt{N_i} m_i(w)\Delta - \frac{1}{\sqrt{N_i}}$

$$A_N f(\underline{z}) = \lim_{\Delta \to 0} [\sum_{i=1}^{K} \{ f(z_1, \ldots, z_i + \frac{1}{\sqrt{\ell_i}\sqrt{N}} - \sqrt{\ell_i}\sqrt{N} m_i'(w)\Delta, \ldots, z_K) \lambda_i(N)\Delta$$

$$+ f(z_1, \ldots, z_i - \frac{1}{\sqrt{\ell_i}\sqrt{N}} - \sqrt{\ell_i}\sqrt{N} m_i'(w)\Delta, \ldots, z_K) \mu_i(N)\Delta \}$$

$$+ f(z_1 - \sqrt{\ell_1}\sqrt{N} m_1'(w)\Delta, \ldots, z_K - \sqrt{\ell_K}\sqrt{N} m_K'(w)\Delta)$$

$$\times (1 - \sum_{i=1}^{K} \{ (\lambda_i(N) + \mu_i(N))\Delta \})$$

$$- f(z_1, z_2, \ldots, z_K) + o(\Delta)] \frac{1}{\Delta} \qquad \text{(A-4)}$$

where for simplicity (and generality) we abbreviate

$$\lambda_i(N) = \lambda_i [N_i - N_i m_i(w) - \sqrt{N_i} z_i] (\sum_{j=1}^{K} (N_j m_j(w) + \sqrt{N_j} z_j)) \qquad \text{(A-5)}$$

$$\sim (\lambda_i N) [\ell_i (1 - m_i(w)) - \frac{\sqrt{\ell_i}}{\sqrt{N}} z_i] (N \sum_{j=1}^{K} (\ell_j m_j(w) + \frac{\sqrt{\ell_j}}{\sqrt{N}} z_j)) \qquad \text{(A-6)}$$

and

$$\mu_i(N) = \mu_i(N_i m_i(w) + \sqrt{N} z_i) \sim \mu_i N(\ell_i m_i(w) + \frac{\sqrt{\ell_i}}{\sqrt{N}} z_i) . \qquad \text{(A-7)}$$

Upon passage to the limit via Taylor series expansion it is seen that

$$A_N f(\underline{z}) = \sum_{i=1}^{K} \{ f(z_1, \ldots z_i + \frac{1}{\sqrt{\ell_i}\sqrt{N}}, \ldots, z_K) \lambda_i(N)$$

$$+ f(z_1, \ldots, z_i - \frac{1}{\sqrt{\ell_i}\sqrt{N}}, \ldots, z_K) \mu_i(N) \}$$

$$+ f(z_1, \ldots z_i, \ldots, z_K) [- \sum_{j=1}^{K} (\lambda_j(N) + \mu_j(N))]$$

$$- \sum_{j=1}^{K} f'_{z_j} \sqrt{\ell_j}\sqrt{N}\, m'_j(w) \quad . \qquad \text{(A-8)}$$

Note that no specific normalization has been utilized up to this point. Now, however, invoke the HTN of (4.6) and utilize (A-6) and (A-7) specifically; allowing N to become large,

$$A_N f(z) \sim \sum_{i=1}^{K} \{ f'_{z_i} \frac{\lambda_i(N) - \mu_i(N)}{\sqrt{\ell_i}\sqrt{N}} + \frac{1}{2} f''(z_i) \frac{\lambda_i(N) + \mu_i(N)}{\ell_i N}$$

$$- f'_{z_i} \sqrt{\ell_i}\sqrt{N}\, m'_i(w) \} \qquad \text{(A-9)}$$

$$\sim \sum_{i=1}^{K} f'_{z_i} \{ \sqrt{N} [\lambda'_i \sqrt{\ell_i} (1 - m_i(w)) \sum_{j=1}^{K} \ell_j m_j(w) - \mu_i \sqrt{\ell_i} m_i(w)$$

$$- \sqrt{\ell_i} m'_i(w)] + \lambda'_i \sqrt{\ell_i} (1 - m_i(w)) (\sum_{j=1}^{K} \sqrt{\ell_j} z_j) - \lambda'_i z_i \sum_{j=1}^{K} \ell_j m_j(w)$$

$$- \mu_i z_i + O(N^{-1/2}) \} + \frac{1}{2} \sum_{i=1}^{K} f''_{z_i} \{ \lambda'_i (1 - m_i(w))$$

$$\times \sum_{j=1}^{K} \ell_j m_j(w) + \mu_i m_i(w) + O(N^{-1/2}) \} .$$

Choose the functions m_i so that they satisfy the system of differential equations (4.5). Let $N \to \infty$. Then for f in the above class of functions, the operator A_N converges to yield

$$A_\infty f(z) = \sum_{i=1}^{K} f'_{z_i} \{\lambda_i' \sqrt{\ell_i}(1-m_i(w)) \sum_{j=1}^{K} \sqrt{\ell_j} z_j - \lambda_i' z_i \sum_{j=1}^{K} \ell_j m_j(w) - \mu_i z_i\}$$

$$+ \frac{1}{2} \sum_{i=1}^{K} f''_{z_i} \{\lambda_i'(1-m_i(w))(\sum_{j=1}^{K} \ell_j m_j(w)) + \mu_i m_i(w)\}$$

The operator A_∞ is the infinitesimal operator of the diffusion whose stochastic differential equation is (4.6) [cf. Arnold (1974), page 152]. The Trotter-Kato theorem can now be applied to assert that the semigroups converge [cf. Burman (1979)].

REFERENCES

Arnold, L. (1974). Stochastic Differential Equations: Theory and Applications, John Wiley and Sons, New York.

Burman, D. (1979). An Analytic Approach to Diffusion Approximations in Queueing. Unpublished Ph.D. Dissertation, Department of Mathematics, New York University.

Coddington, E.A. and N. Levinson. (1955). Theory of Ordinary Differential Equations, McGraw-Hill, New York.

Coffman, E.G., R.R. Muntz, and H. Trotter. (1970). "Waiting Time Distributions for Processor-sharing Systems," J. Assn. for Comp. Mach. 17, pp. 123-130.

Dynkin, E.B. (1965). Markov Processes, Vol. 1, Springer-Verlag, Berlin.

Feller, W. (1957). An Introduction to Probability Theory and Its Applications, Vol. I, 2nd Edition, John Wiley and Sons, New York.

Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. 2, 2nd Edition, John Wiley and Sons, New York.

Gaver, D.P., P.A. Jacobs, G. Latouche (1984). "The Normal Approximation and Queue Control for Response Times in a Processor-shared Computer System Model." Naval Postgraduate School Technical Report, NPS 55-84-001.

Iglehart, D.L. (1965). "Limiting Diffusion Approximations for the Many-server Queue and Repairman Problem," J. Appl. Prob. 2, pp. 429-441.

Kato, T. (1976). Perturbation Theory for Linear Operators, Springer-Verlag, Berlin.

Kelly, F.P. (1979). Reversibility and Stochastic Networks, John Wiley and Sons, New York.

Lehoczky, J.P. and Gaver, D.P. (1981). "Diffusion Approximations for the Cooperative Service of Voice and Data Messages," J. Appl. Prob. 18, pp. 660-671.

Lewis, P.A.W. and L. Uribe (1981). "The New Naval Postgraduate School Random Number Package--LLRANDOMII," Naval Postgraduate School Technical Report, NPS-81-005.

Mitra, D. (1981). Waiting Time Distributions from Closed Queueing Network Models of Shared Processor Systems, Bell Laboratories Report.

Mitra, D. and J.A. Morrison (1983). "Asymptotic Expansions of Moments of the Waiting Time in Closed and Open Processor-sharing Systems with Multiple Job Classes," Adv. Appl. Prob. 15 (1983), 813-839.

Morrison, J.A. and D. Mitra. "Heavy-usage Asymptotic Expansions for the Waiting Time in Closed Processor-sharing Systems with Multiple Classes." To appear in Adv. Appl. Prob., March 1985.

Ott, T.J. (1984). "The Sojourn Time Distribution in the M/G/1 Queue with Processor Sharing," J. Appl. Prob. 21, pp. 360-378.

Pornsuriya, S. (1984). Normal Approximations for Response Time in a Processor-shared Computer System Model, M.S. Thesis, Naval Postgraduate School, Monterey, California.

Ramaswami, V. (1984). "The Sojourn Time in the GI/M/1 Queue with Processor Sharing," J. Appl. Prob. 21, pp. 437-442.

Trotter, H.F. (1974). "Approximation and Perturbation of Semigroups." Butzer, P.L. and Szokefalvi-Nagy, B., Ed., Conference on Linear Operators and Approximation, 2nd, 1974, Mathematisches Forshungsinstitut Oberwolfach Proceedings, Basel, Birkhauser.

NO. OF COPIES

Dr. D. F. Daley
Statistics Dept. (I.A.S)
Australian National University
Canberra A.C.T. 2606
AUSTRALIA — 1

Prof. F. J. Anscombe
Department of Statistics
Yale University, Box 2179
New Haven, CT 06520 — 1

Dr. David Brillinger
Statistics Department
University of California
Berkeley, CA 94720 — 1

Dr. R. Gnanadesikan
Bell Telephone Lab
Murray Hill, NJ 07733 — 1

Prof. Bernard Harris
Dept. of Statistics
University of Wisconsin
610 Walnut Street
Madison, WI 53706 — 1

Dr. D. R. Cox
Department of Mathematics
Imperial College
London SW7
ENGLAND — 1

A. J. Lawrance
Dept. of Mathematical Statistics
University of Birmingham
P. O. Box 363
Birmingham B15 2TT
ENGLAND — 1

Professor W. M. Hinich
University of Texas
Austin, TX 78712 — 1

Dr. John Copas
Dept. of Mathematic Statistics
University of Birmingham
P. O. Box 363
Birmingham B15 2TT
ENGLAND — 1

P. Heidelberger
IBM Research Laboratory
Yorktown Heights
New York, NY 10598 1

Prof. M. Leadbetter
Department of Statistics
University of North Carolina
Chapel Hill, NC 27514 1

Dr. D. L. Iglehart
Dept. of Operations Research
Stanford University
Stanford, CA 94350 1

Dr. D. Vere Jones
Dept. of Mathematics
Victoria University of Wellington
P. O. Box 196
Wellington
NEW ZELAND 1

Prof. J. B. Kadane
Dept. of Statistics
Carnegie-Mellon
Pittsburgh, PA 15213 1

J. Lehoczky
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213 1

Dr. J. Maar (R51)
National Security Agency
Fort Meade, MD 20755 1

Dr. M. Mazumdar
Dept. of Industrial Engineering
University of Pittsburgh
Oakland
Pittsburgh, PA 15235 1

Prof. M. Rosenblatt
Department of Mathematics
University of California-San Diego
La Jolla, CA 92093 1

Prof. Rupert G. Miller, Jr.
Statistics Department
Sequoia Hall
Stanford University
Stanford, CA 94305 1

Prof. I. R. Savage
Dept. of Statistics
Yale University
New Haven, CT 06520 1

Paul Shaman
National Science foundation
Mathematical Sciences Section
1800 G. Street, NW
Washington, DC 20550 1

Prof. W. R. Schucany
Dept. of Statistics
Southern Methodist University
Dallas, TX 75222 1

Prof. D. C. Siegmund
Dept. of Statistics
Sequoia Hall
Stanford University
Stanford, CA 94305 1

Prof. H. Solomon
Department of Statistics
Sequoia Hall
Stanford University
Stanford CA 94305 1

Dr. Ed Wegman
Statistics & Probability Program
Code 411(SP)
Office of Naval Research
Arlington, VA 22217 1

P. Welch
IBM Research Laboratory
Yorktown Heights, N. Y. 10598 1

Dr. Marvin Moss
Office of Naval Research
Arlington, VA 22217 1

Dr. Roy Welsh
Sloan School
M. I. T.
Cambridge, MASS 02139 1

Pat Welsh
Head, Polar Oceanography Branch
Code 332
Naval Ocean Research & Dev. Activity
NSTL Station, MS 39529 1

Dr. Douglas de Priest
Statistics & Probability Program
Code 411(SP)
Office of Naval Research
Arlington, VA 22217 — 1

Dr. Morris DeGroot
Statistics Department
Carnegie-Mellon University
Pittsburgh, PA 15235 — 1

Prof. J. R. Thompson
Dept. of Mathematical Science
Rice University
Houston, TX 77001 — 1

Prof. J. W. Tukey
Statistics Department
Princeton University
Princeton, NJ 08540 — 1

Daniel H. Wagner
Station Square One
Paoli, PA 19301 — 1

Dr. Colin Mallows
Bell Telephone Laboratories
Murray Hill, NJ 07974 — 1

Dr. D. Pregibon
Bell Telephone Laboratories
Murray Hill, NJ 07974 — 1

Dr. Jon Kettenring
Bell Core
435 So. St.
Morris Township, NJ 07960 — 1

Dr. David L. Wallace
Statistics Dept.
University of Chicago
5734 S. University Ave.
Chicago, ILL. 60637 — 1

Dr. F. Mosteller
Dept. of Statistics
Harvard University
Cambridge, MA 02138 — 1

Prof. John B. Copas
Dept. of Statistics
University of Birmingham
P. O. Box 363
Birmingham B15 2TT
ENGLAND — 1

Prof. Donald P. Gaver
Code 55Gv
Naval Postgraduate School
Monterey, California 93943 — 15

Assoc. Prof. Patricia Jacobs
Code 55Jc
Naval Postgraduate School
Monterey, California 93943 — 10

Dr. Guy Fayolle
I.N.R.I.A.
Dom de Voluceau-Rocquencourt
78150 Le Chesnay Cedex
FRANCE — 1

Dr. M. J. Fischer
Defense Communications Agency
1860 Wiehle Avenue
Reston, VA 22070 — 1

Prof. George S. Fishman
Cur. in OR & Systems Analysis
University of North Carolina
Chapel Hill, NC 20742 — 1

Prof. D. L. Iglehart
Department of Operations Research
Stanford University
Stanford, CA 94350 — 1

Prof. Guy Latouche
University Libre Bruxelles
C. P. 212
Blvd De Triomphe
B-1050 Bruxelles
BELGIUM — 1

Library
Code 1424
Naval Postgraduate School
Monterey, CA 93943 — 4

Dr. Alan F. Petty
Code 7930
Naval Research Laboratory
Washington, DC 20375 — 1